W152–W156 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 4 June 2012*
doi:10.1093/nar/gks526

MAGNET: MicroArray Gene expression and Network Evaluation Toolkit

George C. Linderman[1], Mark R. Chance[2] and Gurkan Bebek[2,3,*]

[1]Department of Biomedical Engineering, [2]Center for Proteomics and Bioinformatics, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106 and [3]Genomic Medicine Institute, Cleveland Clinic, 9950 Euclid Ave, Cleveland, OH 44195, USA

Received March 3, 2012; Revised May 9, 2012; Accepted May 11, 2012

ABSTRACT

MicroArray Gene expression and Network Evaluation Toolkit (MAGNET) is a web-based application that provides tools to generate and score both protein–protein interaction networks and coexpression networks. MAGNET integrates user-provided experimental measurements with high-throughput proteomic datasets, generating weighted gene–gene and protein–protein interaction networks. MAGNET allows users to weight edges of protein–protein interaction networks using a logistic regression model integrating tissue-specific gene expression data, sub-cellular localization data, co-clustering of interacting proteins and the number of observations of the interaction. This provides a way to quantitatively measure the plausibility of interactions in protein–protein interaction networks given protein/gene expression measurements. Secondly, MAGNET generates filtered coexpression networks, where genes are represented as nodes, and their correlations are represented with edges. Overall, MAGNET provides researchers with a new framework with which to analyze and generate gene–gene and protein–protein interaction networks, based on both the user's own data and publicly available –omics datasets. The freely available service and documentation can be accessed at http://gurkan.case.edu/software or http://magnet.case.edu.

INTRODUCTION

A large amount of protein–protein interactions (PPIs) and gene expression data has become recently available from high-throughput techniques, such as yeast two-hybrid arrays, microarray gene expression arrays and whole transcriptome shotgun sequencing. Known PPIs are often collected into publicly available databases such as IntAct (1), BioGrid (2) and human protein reference database (HPRD) (3). It is natural to model both protein–protein and gene–gene interactions as a graph, where nodes correspond to genes/proteins, and edges correspond to interactions. Modeling protein/gene interactions as a network allows researchers to use a systems perspective in studying the relationships between different genes/proteins and allows for a host of new analysis techniques. These techniques include generating coexpression networks from mRNA gene expression data (4), PPI networks (5), gene regulatory networks (6), signaling pathways (7), probabilistic networks (8), and predicting reference networks by integrating datasets (9).

Earlier, methods integrating heterogeneous types of high-throughput biological data were presented for gene function prediction (10), biological network discovery (11) and comparative interaction network analysis (12). Species-specific data mining and integration tools/portals also have been developed for *Arabidopsis thaliana* (13), *Drosophila melanogaster* (14), and *Saccharomyces cerevisiae* (15). However, the interactomes generated in recent years using high-throughput data have limited specificity, and the noisy and incomplete nature of the data undermines the results in many promising studies (16). We present an easy-to-use online toolbox, the MicroArray Gene expression and Network Evaluation Toolkit (MAGNET) that provides a solution to this problem. MAGNET integrates publicly available –omics data and user-provided gene/protein expression data into a logistic regression model to provide a weighted PPI, corresponding with the probability that it is a true interaction (17). In addition to weighting PPI networks, MAGNET can generate coexpression networks of user-defined sets of genes using corresponding mRNA expression data, where the associated weights correspond to the Pearson's or Spearman's Correlation Coefficients.

MAGNET's web interface was developed to accept individual experiments from the largest public repository for high-throughput gene expression data, Gene Expression Omnibus (GEO) (18). Users can download files from GEO and directly submit them to MAGNET. Although

*To whom correspondence should be addressed. Tel: +1 216 368 4541; Fax: +1 216 368 6846; Email: gurkan.bebek@case.edu

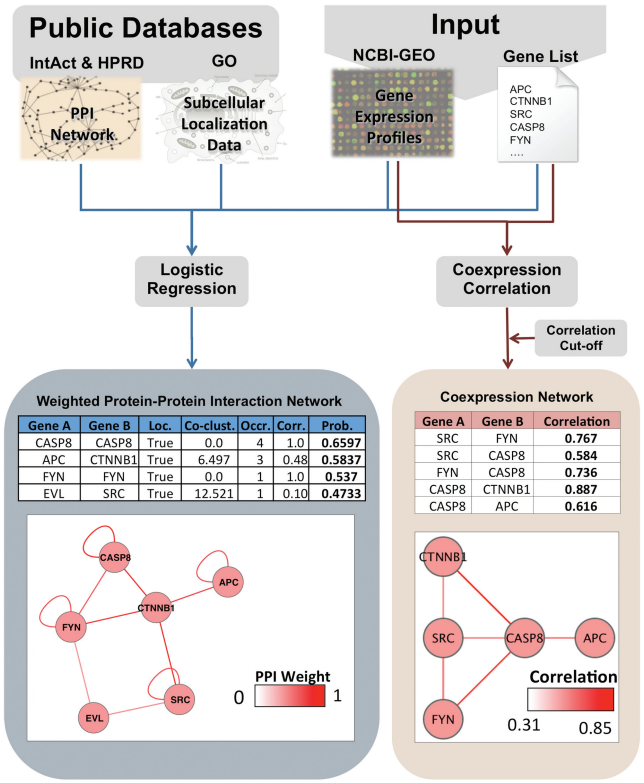

Figure 1. MAGNET processes are shown. The user is asked to supply gene expression datasets and gene list(s) as necessary. The integrated databases are shown at top, and the final output for each process is shown in boxes below. The networks can be viewed as both tables and interactive graphs drawn with a web-based network viewer.

GEO is easily accessible, it is often cumbersome to read, filter and analyze these files as most exceed the capabilities of modern spreadsheet software. When weighting PPI networks, the user simply must specify the types of publicly available data to incorporate into the logistic regression model (localization, literature references, co-clustering), and the toolbox retrieves the data from its own database. This allows researchers to harness the power of a diverse group of public databases without having to deal with each of the different formats and standards (Figure 1). Overall, MAGNET provides value by easily processing and visualizing system-wide datasets to the end of generating or prioritizing interactions for further evaluation.

MAGNET WEB SERVER

Scoring protein–protein interaction networks

MAGNET assigns weights to known PPIs by integrating sub-cellular localization data, co-clustering coefficient, number of literature observations and user-provided mRNA-level co-expression data (17). Each of these four variables, $\vec{X} = (x_1,x_2,x_3,x_4)$, is incorporated into a logistic regression model, where the probability of a true interaction between two proteins i and j given the variables $\vec{X}_{ij}$ is $\Pr(I_{ij} = true|\vec{X}_{ij}) = \frac{1}{1+\exp(-\beta_0-\Sigma_{i=1}^{4} x_i\beta_i)}$ (17,19). Using a 'golden' dataset of experimentally verified interactions as a positive training set (20), and randomly selected (non-golden) interactions as negative training set (500 each), the model is trained to the specific experiment. By re-training MAGNET for every job that is submitted, MAGNET can find the optimal coefficients for each of the variables depending on its usefulness in determining the plausibility of the given interaction. MAGNET repeats the training step for a user-defined number of iterations and then takes an average over the resulting variable weights to determine the final set of constants $(\beta_0,...,\beta_4)$. For example, in some cases localization data were more useful and hence given a higher weight, whereas in others the localization data were not as beneficial. MAGNET uses this trained logistic model to score the PPI network, generating a weighted network. Therefore, each edge is associated with a probability showing how likely it is that the interaction exists, given the values of the four variables. It is important to note that the model is trained for every job that is submitted, which results in a model that is trained specifically to score interactions based on the user-provided microarray data.

The first of the four variables, sub-cellular localization, is based upon the reasonable assumption that two interacting proteins are more likely to interact if they are co-localized to the same cellular component. The localization information is obtained from Gene Ontology *cellular component* annotations (21). A positive value (+1) is assigned if the proteins share at least one sub-compartment, whereas a negative value (−1) is assigned if they do not. While most of the proteins have this information, if there is no annotation found, they are scored with zero (0) to avoid unnecessarily penalizing these interactions. The second variable, the co-clustering coefficient, measures the connectedness of the neighbors of two given proteins, which has been shown to suggest a higher probability of interaction (22). The third variable measures the number of times that a given interaction has been reported across the PPI databases. The fourth variable is the correlation coefficient (Pearson's or Spearman's) between the expression values of the two genes corresponding to the given proteins. These correlations are calculated based on the user-provided expression data, whereas the former three variables are integrated into MAGNET. Expression data is the exception because expression of an interacting pair may vary greatly depending on the samples chosen.

By integrating the first three variables from public databases with the fourth variable from tissue-specific-gene/protein expression measurements, MAGNET effectively allows researchers to harness the power of these datasets while still obtaining results specific to their experiment. Suthram *et al.* (23) have evaluated various models used to assess the quality of interaction confidence assignment schemes. It was reported that a similar logistic regression model (without co-localization) performs better than others in correlating functional assignments of proteins. Hence, MAGNET utilizes a logistic regression model, since our focus is on assessing the validity of functional relationships of protein pairs in a given system.

Web Server

To submit a job to score a PPI network, the user must upload normalized expression data and the platform definition files (available from GEO) describing the genes targeted in that platform. The user can also select the variables used in the logistic regression model (Figure 2). After submitting the expression data, the user can filter the samples by the 'sample_characteristic_ch' fields in the GSE file (Step 2). This allows the user to include or exclude specific samples without having to manually edit the files. Finally, the user is presented with the console output and can then proceed to the Results page, where the resulting PPI network can be viewed as a network with a web-based network viewer, an edge list table, or downloaded as Cytoscape readable files for further analysis.

Generating coexpression networks

Method

Analysis of coexpression relationships between genes/proteins provides insights into their interactions and functions. MAGNET can quickly and easily generate coexpression networks for a given set of genes where genes are represented by nodes, and edges connect genes whose coexpression correlation (Pearson's or Spearman's) is above a certain cut-off value (Figure 3). Although similar tools are available, they require specific platforms (e.g. R or Matlab) and do not provide easy access to visualization tools (4,24). This module works independently of the PPI module, and it generates networks that quantify the pair-wise correlation of the genes in a given network, i.e., high correlation values reflecting coexpression and negative correlations reflecting differential expression.

Web Server

The form to generate a coexpression network is similar to that of the PPI network module, except that the user can also specify cut-offs to filter the edges in the resulting coexpression network. After job submission, the user is presented with the console output and can then proceed to a Results page similar to that of the PPI network module.

SOFTWARE DOCUMENTATION

MAGNET provides both an online manual and a full tutorial. In this manual, the workflow of MAGNET is explained step-by-step. Additional pictures and screenshots can guide the user who wants to understand the details and to tune the parameters available for MAGNET users. For testing purposes only, MAGNET provides an exemplary expression data file that can easily be used by selecting sample data during the first step of the wizard in each module. By doing so, users find a quick way to examine MAGNET's features. The data provided are publicly available from GEO (GSE 19338 [19]) and represent gene expression profiling experiments run from villus and crypt layers of murine intestine. The series includes profiles from wild-type mice and mice that have mutations in adenomatous polyposis coli (APC) and p21 genes. These data were normalized using robust multi-array (RMA) normalization and uploaded with detailed annotations for future filtering steps. We use a gene set of 11 genes as an example to test the two processes, although

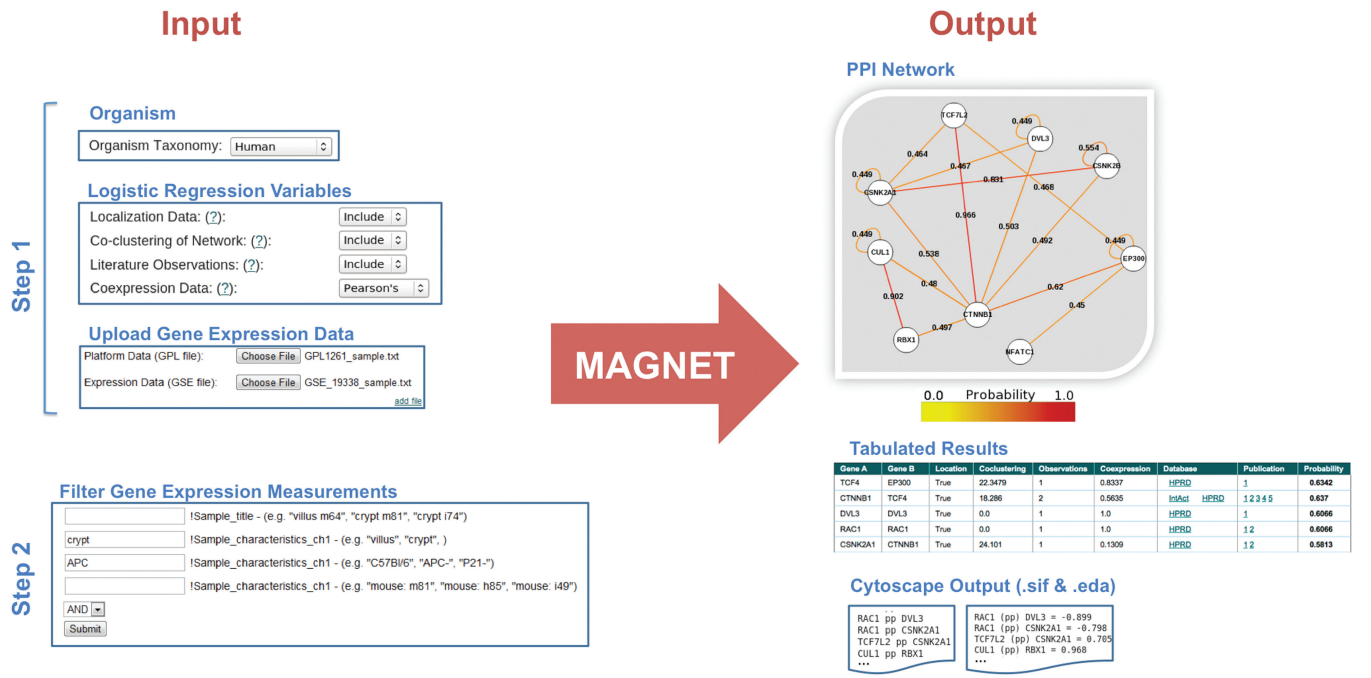

Figure 2. Workflow for Module 1: Weighting protein–protein interaction networks. In step 1, the user specifies the organism taxonomy, adjusts the variables included in the logistic regression model, and uploads the gene expression data in GEO-compatible format. In step 2, the user can filter the samples based on the sample characteristics and annotation. After processing, the weighted PPI network is available in Cytoscape-compatible format, an interactive web-based network viewer for visualization and as a browser table with links to external source databases.

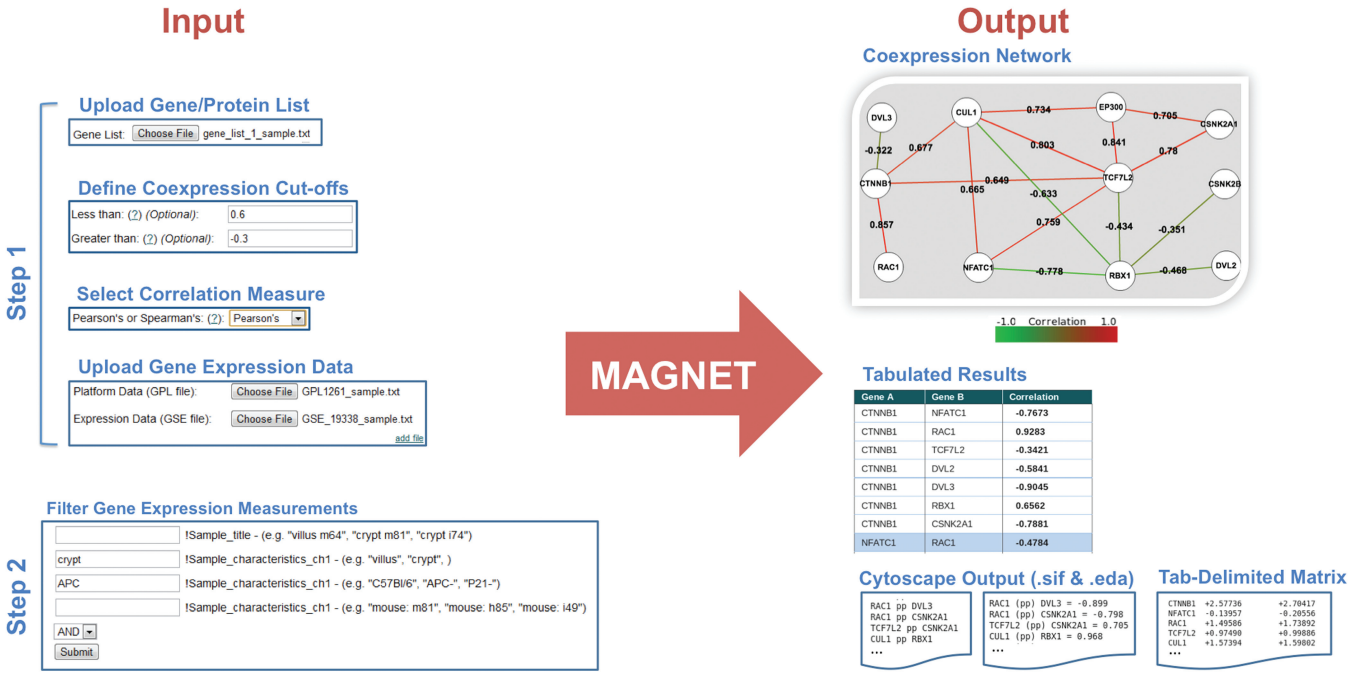

Figure 3. Workflow for Module 2: Generation of coexpression networks. In step 1, the user uploads selected genes (or leaves it blank if interested in all genes in the array), specifies the coexpression cutoffs (if filtering is desired), specifies the type of correlation measure, and uploads the gene expression data in GEO-compatible format. Then, the user is taken to step 2, where the samples can be filtered by their annotation and sample characteristics. After processing, the output consists of tabulated results, a spreadsheet of correlation values, Cytoscape .sif and .eda files, and an interactive web-based network viewer for visualization.

MAGNET does not limit the number of genes in a given job.

CONCLUSION

MAGNET allows users to both score PPI networks by integrating four different diverse data types and to generate coexpression networks given expression profiles. All modules are developed for expression data formatted in the GEO SOFT format, but the site contains templates for non-GEO data as well. The site is optimized to work with large datasets with ease, preventing the user from having to deal with cumbersome arrays prior to analysis. The tool is freely available to researchers and can be accessed with any up-to-date web browser available.

ACKNOWLEDGEMENTS

The authors thank Shannon Swiatkowski for coming up with the acronym MAGNET.

FUNDING

National Institutes of Health (NIH) [P30-CA043703 to M.R.C. and UL1-, BB123456 to M.R.C.]. Funding for open access charge: NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–531.
2. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–704.
3. Prasad,T.S.K., Kandasamy,K. and Pandey,A. (2009) Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.*, **577**, 67–79.
4. Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article17.
5. Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F. and Gerstein,M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
6. Margolin,A.A., Nemenman,I., Basso,K., Wiggins,C., Stolovitzky,G., Dalla Favera,R. and Califano,A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinfo.*, **7(Suppl. 1)**, S7.
7. Bebek,G., Patel,V. and Chance,M.R. (2010) PETALS: Proteomic evaluation and topological analysis of a mutated locus' signaling. *BMC Bioinformatics*, **11**, 596.
8. Lee,I., Date,S.V., Adai,A.T. and Marcotte,E.M. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.
9. Srinivasan,B.S., Shah,N.H., Flannick,J.A., Abeliuk,E., Novak,A.F. and Batzoglou,S. (2007) Current progress in network research: toward reference networks for key model organisms. *Brief Bioinform.*, **8**, 318–332.

10. Troyanskaya,O.G., Dolinski,K., Owen,A.B., Altman,R.B. and Botstein,D. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci. USA.*, **100**, 8348–8353.
11. Myers,C.L., Robson,D., Wible,A., Hibbs,M.A., Chiriac,C., Theesfeld,C.L., Dolinski,K. and Troyanskaya,O.G. (2005) Discovery of biological networks from diverse functional genomic data. *Genome Biol.*, **6**, R114.
12. Alexeyenko,A., Schmitt,T., Tjarnberg,A., Guala,D., Frings,O. and Sonnhammer,E.L. (2012) Comparative interactomics with Funcoup 2.0. *Nucleic Acids Res.*, **40**, D821–828.
13. De Bodt,S., Carvajal,D., Hollunder,J., Van den Cruyce,J., Movahedi,S. and Inze,D. (2010) CORNET: a user-friendly tool for data mining and integration. *Plant Physiol.*, **152**, 1167–1179.
14. McQuilton,P., St Pierre,S.E. and Thurmond,J. (2012) FlyBase 101—the basics of navigating FlyBase. *Nucleic Acids Res.*, **40**, D706–714.
15. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–705.
16. Stumpf,M.P., Thorne,T., de Silva,E., Stewart,R., An,H.J., Lappe,M. and Wiuf,C. (2008) Estimating the size of the human interactome. *Proc. Natl Acad. Sci. USA.*, **105**, 6959–6964.
17. Bebek,G. and Yang,J. (2007) PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics*, **8**, 335.
18. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–890.
19. Patel,V.N., Bebek,G., Mariadason,J.M., Wang,D., Augenlicht,L.H. and Chance,M.R. (2010) Prediction and testing of biological networks underlying intestinal cancer. *PLoS One*, **5**, e12497.
20. Mewes,H.W., Heumann,K., Kaps,A., Mayer,K., Pfeiffer,F., Stocker,S. and Frishman,D. (1999) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **27**, 44–48.
21. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**, 25–29.
22. Goldberg,D.S. and Roth,F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci. USA.*, **100**, 4372–4376.
23. Suthram,S., Shlomi,T., Ruppin,E., Sharan,R. and Ideker,T. (2006) A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, **7**, 360.
24. Ruan,J., Dean,A.K. and Zhang,W. (2010) A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC Systems Biol.*, **4**, 8.
25. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.